PRE-POST OUTCOME ANALYSES:

Webinar for Teen Pregnancy Prevention 2018 Grantees

Friday, February 28, 2020

1:00 – 2:00 p.m. Eastern Time

Webinar transcript

Webinar producer: Hello, everyone, and thank you for attending today’s event. Before we begin, we want to cover a few housekeeping items. At the bottom of your audience console is a multiple application which you can use. You can expand windows in the console by clicking on the maximize icon in the top right of the widget or by dragging the bottom right corner of the widget. If you have any questions during the webcast, you can click on the Q&A widget at the bottom and submit your question. We will try to answer these during the webcast, but if a fuller answer is needed or we run out of time, it will be answered later via email. We do capture all questions. If you have any technical difficulties, please click on the help widget. It has a question mark icon and covers technical issues. However, you can also submit technical questions through the Q&A widget. An on-demand version of the webcast will be available approximately one day after the webcast and can be accessed using the same audience link that was sent to you earlier. The recording and materials will be posted next week to the max.gov website. Now I’d like to turn it over to Diana McCallum. Diana, you now have the floor.

Diana McCallum: Hi, everyone. Welcome to today’s webinar. As you've seen from our emails, we have two upcoming webinars that are really designed to help you strengthen the journal articles that you'll be developing over the coming months. Today, Russell Cole will discuss pre-post outcome analyses and the webinar on March 9 is going to focus on conducting qualitative analysis. Many of you may know Russell Cole is a senior researcher here at Mathematica with expertise in research design, evaluation technical assistance, and systematic reviews. He has really extensive experience providing evaluation TA [technical assistance] to teen pregnancy prevention [TPP] grantees. He’s been doing so for almost a decade. As a principal investigator and deputy project director on a previous evaluation TA contract, he provided oversight, to teams of researchers and oversees technical assistance for more than 40 local TPP grantee evaluations for 2010 grantees and more than 20 local evaluations for 2015 grantees.

So, based on your journal article abstracts and our discussions with you during TA calls, we know that many of you are planning pre-post analyses. So we hope that this webinar gives you an opportunity to think about strategies that you're going to use to help you strengthen your analyses, and help you think through any questions you might want to discuss with us today or that you might want to pose later to the Evaluation TA team. During the webinar, we invite you to submit your questions through the Q&A widget. And we'll respond to them at the end right. And now I'll turn the presentation over to Russ.

Russell Cole: Thanks Diana. That was really nice introduction. So, yes, I'm Russ Cole. And today we're going to be talking about pre-post outcome analyses, notably talking about how they can be used to establish the foundation of an argument that your program is ready for an impact evaluation.

Agenda

So, today's presentation is structured around three topics. The first is presenting the value of a pre-post study. What's the purpose of one? What does it get you? Also what doesn't it get you? And really how does a pre-post study set the stage for a future impact evaluation. The second topic is doing a pre-post study. These are the basics. How does one do this type of analysis? What kinds of statistics are important to report? And the third topic is additional analyses. This is going to be the bulk of the presentation. It's really going beyond the basics to make your pre-post findings more credible. So my goal is to finish today in about 45 minutes. That'll leave plenty of time for some Q&A at the end.

Evidence from a pre-post outcome study

What is a pre-post outcome study?

Let's begin with the first piece, the types of evidence we get from a pre-post outcome study. Let's start with some basic definitions shown here. A pre-post study is one that quantifies how participants' outcomes change over the course of a study and typically compares how participants' outcomes change between a baseline or pre-intervention period to program exit or a follow-up period—that difference in participants' outcomes from baseline or pre to follow-up or post represents individual change. And by aggregating or averaging this difference across all program participants, we can quantify how outcomes changed on average.

We're going to primarily focus today on change analyses for two assessment points, but the approaches shown here could be used for longer-term follow-ups to assess baseline to long-term results. If you're in that boat, the approach is no different. Also, some outcome analyses are going to have multiple time points; a common approach is to fit a trend line through the data. The same general ideas are true for longitudinal analysis. But we're going to focus on pre- and post-[analysis] based on two assessment points because everyone's going to have those data and understanding those principles establishes the foundation for more complex analyses or three or more time points.

How to interpret average pre-post outcome change?

Alright, so what is average change? It's how individual outcomes among program participants change over time on average. Importantly, average change is not equal to the impact or the effect of the program. Importantly, we cannot attribute the change in outcomes to the program being tested in these types of analyses. And the key is that without a counterfactual, we cannot disentangle changes in outcomes that are solely attributable to the program that we're testing from naturally occurring changes in outcomes. So the types of changes that would occur naturally through maturation, testing, or regression. These are the kinds of things that Campbell and Stanley talked about in their experimental and quasi-experimental designs for research texts from 1975.

So, therefore, it's really important to accurately describe these findings and their limitations appropriately. I've got some illustrative text here to point out how to talk about these types of findings. For example, you could say between program entry and program exit, participants' knowledge scores improved by 30 percentage points. That's the pre-post outcome finding, but here's the caveat language. This analysis assesses individual change over time without a counterfactual. It's not appropriate to assert that the program was solely responsible for the observed improvement in outcomes.

How can pre-post findings create the foundation for an impact study?

So, at this point, we know what a pre-post outcome study is and how to appropriately interpret findings. The question now is how to frame these findings as establishing the foundation for why an impact study is the next logical step in evidence building. It really comes down to the two questions that are shown here. First, are outcomes moving in the right direction? Your logic model or your theory of change for an intervention presents a hypothesis about which outcomes might change and when they might change. So these pre-post outcome findings provide the data to test that hypothesis. At a minimum, the most proximal outcomes in the logic model or the theory of change should change or should improve over time.

Secondly, are the changes in the outcomes large in magnitude for many outcomes? We might expect to see improvement in the absence of a program. What I'm going to argue is that these changes in outcomes might represent an upper-bound estimate of program effectiveness for a future impact evaluation. When we're doing an impact evaluation, we commonly do something called a power analysis or statistical power analysis. Given a certain size sample, how big do the impacts have to be for us to be likely to state the difference is statistically significant? What I'm going to argue over in the next few slides is that the pre-post change that we observe is going to give us an upper-bound estimate and optimistic estimate of the types of changes that we might observe in an impact study.

Example: Program participants improved sexually transmitted infection (STI) knowledge by 30 percentage points (30PP)!

So here's an illustrious finding from a typical pre-post outcome study. At baseline, average scores on an STI [sexually transmitted infection] knowledge test were 60 percent. At exit average scores were 90 percent. So that's great. We saw a 30-percentage point increase. It shows improvement in the key outcome that was targeted by the intervention. So outcomes are trending in the right direction. That's a good thing. What about the magnitude? We see a big 30 percentage point increase. That's also great. It's a large change.

So the question then becomes, is this what we should expect to see as the effect of the program, if we were to do this in a future impact study? That is, would the difference in outcomes across the treatment and comparison groups be about 30 percentage points? I'm going to say, no.

Use pre-post impact estimate as upper bound for statistical power calculations

And this slide helps to illustrate why. If you look at this slide, we've got the same information for the treatment group, but now I'm also bringing in a hypothetical comparison group. So this hypothetical comparison groups started at the same place as the treatment group. It had a 60 percent score on the STI knowledge test and it increased to about 70 percent at the follow-up assessment. So it's pretty common for a business-as-usual group to improve. It's not like they're going to get nothing. Plus there's often just natural maturation or growth that's happening among the comparison group.

So again, seeing improvements among the comparison group should be expected. So this is to say, if we did an impact study, the observed impact would only be 20 percentage points here. It's the difference in outcomes that follow-up among the treatment and comparison groups. So this is really a smaller difference than the 30 percentage point improvement we saw in the previous pre-post study. So that's why I argue that the pre-post study offers an upper-bound estimate for what you can expect to see when you're powering for a future impact evaluation. Notably, you should make sure that you're adequately powered to detect impacts that are smaller than those that are observed in your pre- post analysis.

We've established why we do a pre-post analysis, and how they can potentially establish a foundation for what a subsequent impact study might do. Where we’re going to shift gears to now, is the nuts and bolts of actually doing a pre- post analysis and what to report.

Estimating and reporting pre-post differences: The basics

Goal: Describe how individual outcomes change over time

So let's start from an important place. It's critical to choose the right outcomes for a pre-post analysis. Some measures are going to naturally change over time. For example, sexual initiation. The prevalence rates are going to increase as youth get older. So it's going to be a really hard variable to interpret in the absence of a comparison group. You won't be able to tell if an observed change is a good or a bad thing. I'd argue that thinking about knowledge and attitudes, they are probably better outcomes for pre-post [analysis], assuming that they are expected outcomes shown in your logic model or theory of change.

So, with a quick talk about that outcome selection behind us, let's talk about the analytic approaches. First, I recommend conducting within-individual analysis. It means that we're going to match pre and post outcomes for each individual. So this means that we're going to eliminate individuals from the analysis who are missing one or both assessments. Importantly, we're going to do this analysis separately for each outcome of interest. That is, each outcome is going to be its own separate analytic sample. And the benefit of this approach of doing within-individual analysis is ease of interpretation. Doing this type with an individual analysis eliminates compositional differences or biases that can occur if analysis is conducted with all available data. An additional benefit is that when we get to doing the inferential tests, it's very basic to do this.

A limitation of this approach, however, is that the complete case sample might not represent the full study sample. That being said, this is what I'm going to attempt to address in the last half-hour of this call.

Different types of respondents observed in a pre-post study

[Displays pie chart of different respondent types]

Let’s understand compositional changes. Let's take our sample and break it out into four categories or types of respondents based on whether they were respondents to each of our two surveys. Each person in our study must fall into one of these groups for a given outcome. We've got nonresponders shown in blue. These are folks who don't respond to either pre or post [survey]. It's a non-negligible chunk in this pie chart. We've got pre-only [survey] respondents in red. These are folks who respond to the pre-test but not the post-test. We've got the post-only [survey] respondents in gray. These are folks who respond to the post-test but not the pre-test. And then we've got the pre and post respondents. This is our complete case sample. They're shown in yellow. This is the bulk of our sample for this example.

Unpacking issue of composition as a source of bias in understanding individual change

Let's talk about the composition of the sample and assessment scores. We've got that pie chart from the previous slide on the left. And we're going to look at the scores of individuals on the right panel shown as a table. There are three columns in the table. First, we've got the type of the respondent plus their relative prevalence in the pie chart. We've got the average scores of these respondents at pre, as well as the average scores of these respondents at post and each row represents scores for a given type of a respondent. So the first row shows nonrespondents. They aren't observed by definition at either pre or post [survey]. So we don't actually have an average score for them in this table. The pre-only respondents averaged a score of 70 at the baseline. But they're not observed at the follow-up. The post-only respondents averaged a 75 on the post-test, but they don't have a baseline. And the complete case sample average is at 60 pre and an 80 at post.

So if we look at all of the data that we have in hand, the average score of all of the data at the pre-test was 61.3. This is the bottom row. It's essentially a weighted average of the complete case sample, the pre and post respondents with the pre-only respondents. Similarly, the average of all of the observed data at post-test is 79.7.

So my question to you is ,when we talk about how scores changed over time, what's the right way to think about this or what's the right information to report? One approach would be to use all observed data. We could say that scores increased from 61.3 to 79.7. Again, that's the bottom row. The problem here is that this muddles individual change in scores with composition change. We can't simply state that this difference represents the growth or improvement in individual scores over time. This is because it includes folks who are observed at pre, but not post or folks who were observed at post but not pre. We don't actually have the data to know how those folks change. But they're contributing to those change statistics. So we're no longer talking about how individuals are changing over time.

I'm going to argue that it's better to focus on the pre and post that complete the case sample one row above. It helps us to mitigate composition change as contributing to the observed change in outcomes, so that we have a cleaner, more interpretable result. We want to talk about how individuals changed. Complete case analyses help us achieve that goal.

What to report among the complete case sample?

What do we want to report to be transparent about this? Specifically, what do we want to report in a pre-post outcome study that's based on a complete case sample? We want to report the pre and post means and standard deviations. We want to talk about the difference in means both in raw units as well as in standard deviation units. And we want to talk about standard deviation units relative to the post-test period. How do you calculate a difference in means and center deviation units? You just divide the raw difference by the standard deviation at post-test. It's as easy as that.

We want to make sure that the audience understands the practical significance of the change. It depends on the outcome: whether it makes more sense to focus on the raw difference or a standardized difference. We want to report a *p*-value. The easiest way to obtain the *p*-value for this type of analysis is to use a paired t-test. This is one of the benefits of doing a complete case analysis. If your sample sizes are really small, for example, less than 30, I'd recommend that you do a nonparametric test to satisfy the normality assumption and get the right *p*-values. It's called the Wilcoxon signed-rank test. Just to note this, when you do these types of analyses, a paired t-test, you're typically going to have huge power. So don't be surprised when you have tiny *p*-values and everything looks like it's a statistically significant improvement.

You might be saying, “Russ, I'm doing my analysis across multiple schools for example. Do I need to worry about clustering?” No, you don't. We're talking about within-individual change. So you're good to go with that regular paired t-test or t-test analog analysis. That being said, having multiple implementation settings does dovetail nicely with the next slide, which is about talking about heterogeneity and outcomes.

Explore and report heterogeneity in outcome change

The previous slide presented information on how to showcase change for a complete case sample for a single outcome. What I want to argue here is how to do this across outcomes to tell more of a story. Now, first thing that you're going to probably want to do is report results for different outcomes. In particular, the most proximal outcomes to the intervention are likely to be the ones to show the biggest changes or the biggest improvements. And similarly, the more distal outcomes to the intervention might show less change. This is a pretty useful type of analysis to help showcase that your logic model, that your theory of change is on target. So this is a good set of things to report and again to tie back to your logic model.

You might want to also see whether changes in outcomes vary by subgroups of interest. Maybe you want to look at demographically defined subgroups. For example, sex, race, or age if that makes sense or, building from the previous slide, looking at changes and outcomes by features of the implementation site. It's probably useful to acknowledge that some subgroups might create ceiling or floor effects—for example, [members of] young samples might not be sexually active. So it wouldn't be surprising to see no changes in observed outcomes for some of those sample members. But in general, this is a good thing to explore.

A different type of pre-post analysis that's also useful to play around with is one where you use treatment receipt as a variable that defines subgroups of interest. You can create an indicator for whether an individual receives the program as intended. That is, if they got a sufficient dose of the program, for example. And we can compare the improvement in outcomes across groups that got the intended dose or not. You might even want to do a between-group inferential test to explore this a little bit. It's probably not a well-powered test and it's certainly wise to talk about this as an exploratory analysis with lots of limitations. But it is potentially useful. There's a lot of other stuff that would go into such an analysis to add to its credibility. But again, these are the types of things that might be useful to begin to explore in studies where you don't have a comparison group.

Additional analyses to enhance pre-post findings

Skeptical readers will be unsatisfied with the basic presentation

So that was the basics. There was nothing terribly new there. But what I'm going to argue is that let's go beyond. Let's start to think about the things that critical readers will point to as threats to the credibility of those pre-post findings, and ways that we can protect ourselves against them. Importantly, I'm not going to be able to talk about the concern of not having a counterfactual condition. At this point, there's really nothing that we can do about that limitation because we don't have comparison data. But what I am going to talk about is just doing a complete case analysis, a complete case pre-post outcomes analysis and stopping there.

The main threat for the credibility of a pre-post outcome study is that the complete case sample isn't representative of the full study sample. We're going to walk through a bunch of analyses that we can conduct and report on that attempt to address this concern. The subbullets shown here are the things that we're going to cover in the remainder of the presentation today.

Step 1: Response rate analysis

Response rate analysis

[Displays four types of respondents: nonresponder at both assessments, pre only, post only, pre and post (complete case)]

So the first item is doing a response rate analysis for each outcome. The first step is pretty straightforward. We talked about this conceptually earlier. For each outcome, we can categorize each individual in the data set as one of these four types and report the prevalence rates of each type. So that's what we're going to do in a quick example.

Respondent type prevalence rates will vary due to item non-response

[Displays table of respondent type prevalence rates, see following table]

Respondent type	Intention to remain abstinent	Recent sexual behavior
Nonresponders	20%	23%
Pre only	8%	9%
Post only	3%	4%
Pre and post (complete case sample)	69%	64%

So this is probably more transparent than what you're used to seeing, but it does help to establish a foundation for subsequent analyses. We're just seeing the extent to which survey and item nonresponse is affecting who's in and who's out of our pre-post outcome analyses.

There are a couple of things to point out here. What we're hoping for is that in our pre and post sample, our complete case respondents shown in the bottom row are the most prevalent category. And that's the case here. We're hoping that the nonrespondents at both time points are relatively infrequent. Because we're not going to have much or potentially any information about who those folks are. One last thing that the idea that the post-only responders are the most infrequent category. That's a pretty typical situation. It's pretty rare for folks to be missing at baseline and then show up for a follow-up assessment, but it does happen. It's just typically infrequent.

So that was easy; that was the first step. It was to try to get a sense of the prevalence of our complete case sample. If you find that you've got 90 percent of your sample being a complete case sample respondent, you kind of have an argument to not bother doing anything else. Because your complete case sample is dominating your overall sample. It's probably pretty representative on its own. However, in most studies, you probably won't have 90 percent complete case sample respondents. And therefore, it's useful to do additional analyses, starting with predicting who is or who isn't in the complete case sample. In other words, doing a nonresponse analysis.

Step 2: Nonresponse analysis

Nonresponse analysis approach

Here are the basics. We're going to try to predict who the complete case respondents are relative to the target population. Essentially, we're going to do a regression analysis to identify factors that help sort our sample into whether they are this type of respondent or not. So here's the wrinkle. What data are we going to use to predict whether someone is a complete case respondent or not? Probably, the best, the most complete data that we have for predicting someone's propensity to be a complete case respondent is the baseline survey.

But we know there will be some folks who don't respond to the baseline survey. So what are we going to do? And it’s really two solutions. The first solution, the easy solution, is that we try to predict whether folks are complete case respondents among those who have baseline data. The second approach is that we try to predict whether folks are complete case respondents among the full study sample from whom we have any data. And we do imputation to fill in missing data on key variables. This second approach is more complicated. Plus, there's plenty to do even under the previous simple case. So just to keep things straightforward, I'm going to stick with a simple approach. We're going to try to generalize our complete case sample respondents back to the baseline survey respondents. The key benefit here is that we really want a baseline measure of the outcome of interest as a predictor. That's why I'm going to argue that this is an appropriate way to go to move forward.

Approach for assessing nonresponse bias (Step 1)

Alright, here are the nuts and bolts of the approach. We're going to try to identify variables that we think are potentially related to whether an individual is going to be a complete case sample respondent. I've listed a bunch of things that are typically available and are likely useful for this exercise. We're likely to have demographics. We will have a baseline assessment of the outcome of interest. We might have site characteristics, if you have a multisite implementation. For example, perhaps different staff at sites or different levels of experience with youth or with the program. Those types of variables could be used as predictors. You might have other baseline variables in your survey that theory or literature suggest might predict survey response. For example, you might have measures of youth motivation or their grit or their persistence. Those are going to be the predictor variables that we're going to use for our nonresponse analysis.

At this point, you're going to create an indicator variable for whether an individual is in the complete case sample for a particular outcome. For example, if our outcome of interest is STI Knowledge, we would create an STI underscore CC indicator. And that CC stands for complete case. A person gets a 1 for that indicator if they're in the complete case sample and a 0 otherwise.

Illustrative dataset

[Displays table of illustrative dataset, see following table]

StudyID	Male	Hispanic	Age	GRIT	STI_Knowledge_Pre	STI_Knowledge_Post	STI_CC
101	1	0	15.5	8	78	79	1
102	1	0	15.7	7	77	82	1
103	1	1	16.1	8	45		0
104	0	1	15.8	9	56	54	1
105	0	1	15.9	10	65	67	1
106	1	1	16	8	91		0
107	1	1	16.5			95	0
108	1	1	16.5			95	0

This is a lot to take in, so let's look at some data to try to illustrate this. This is an illustrative data set of eight individuals. You can see study ID 101 through 108 on the left-hand side. We've got several rows of demographics, excuse me several columns of demographics. We have an STI Knowledge Pre-test and STI Knowledge Post-test. And we have that STI _CC indicator. That's the new variable that we just created. It takes on a value of 1 when the observation has both a pre- and a post- assessment for the STI measure, and it has a 0 otherwise. So for example, if you look at the third row, study ID 103, that person doesn't have a post-test. So their STI_ CC value is 0.

One last thing, if we look at the bottom row, this individual is not going to contribute to the analysis most likely. Most stats packages use something called listwise deletion to get rid of individuals who are missing key variables, either predictor or outcome variables, for something like a regression analysis. So, if we're going to use as predictors in our nonresponse analysis things like grit or the STI Knowledge pre-test, this person is going to be dropped from the analysis because they have missing data for those key variables.

Approach for assessing nonresponse bias (Step 2)

So what are we going to do with those data? First, we're going to regress the complete case indicator on predictors of interest using something like logit or probit regression. If we think about that last set of data that STI_ CC variable would be our dependent variable. We would have predictors like sex, ethnicity, age, grit, baseline, and STI Knowledge. While I didn't include any cluster level variables, we could certainly include them if those are appropriate for analysis and you'd want to cluster your standard errors as appropriate for predictor variables that are measured at the cluster level. You'd want to interpret the raw and standardized beta coefficients, as well as the *p*-values that are output from your probit or logit analysis. The raw betas are more interpretable, but the standardized betas tell you which variables are the most important relative to the others in the model.

And this final bullet kind of summarizes some key takeaways. Here are the kinds of ways to interpret or present the findings from your nonresponse analysis. You could say things like students receiving free or reduced-price lunch were 2.4 times less likely than non-free and reduced-price lunch students to be included in the complete case sample or the complete case sample tended to represent a lower-risk sample. The baseline assessment of the outcome was the single strongest predictor of whether an individual was in the complete case sample. Or if you wanted to talk about a number of variables in one sentence, you could say something like complete case sample members tended to be non-Hispanic and have high levels of self-reported motivation and intended services at schools rather than community settings.

Illustrative SAS Code and Output

[Displays SAS Code and output:

```
Proc logistic data=my data;
Model STI_CC (event=’1’)= male Hispanic age
STI_knowledge_pre_GRIT/link=logit stb;
Run;
```

Displays analysis of maximum likelihood estimates, see following table]

Parameter	DF	Estimate	Standard error	Wald chi-square	Pr > ChiSq	Standardized Estimate
Intercept	1	1.1096	2.8070	0.1562	0.6926	
Male	1	0.1125	0.3780	0.0886	0.7660	0.0310
Hispanic	1	0.7732	0.5226	2.1892	0.1390	0.1768
Age	1	-0.0256	0.1812	0.0199	.8877	-0.0143
STI_knowledge_pre	1	-0.00345	0.00771	0.2004	0.6544	-0.0473
Grit	1	0.4629	0.0923	25.0890	<.0001	0.7487

These are ways of again trying to put in words what comes out of your statistical package regression results. I'm going to show some illustrative SAS code and output to kind of tie this back together. I know folks here are going to use different software packages. So the syntax isn't going to work for everyone, but the concepts should. And again, the output is going to look comparable. At the top panel of this slide, we can see that we're doing logistic regression. Our dependent variable is whether the person is in the complete case sample for the STI outcome. It's being predicted by a bunch of X variables. Those are the things on the right-hand side of the equal sign and asking the software package for standardized beta coefficients.

I snipped a chunk of the output focusing on the raw and standardized betas from the regression model plus the *p*-values. If you look at this output, you can see that being an STI Knowledge complete case sample member is largely determined by the baseline grit score. It's the only statistically significant predictor. It is the largest standardized estimate on that far right-hand side. If you exponentiate the raw grit beta parameter of 0.46, you can see it's highlighted; you would get the odds ratio of 1.6. So for those of you are familiar with interpreting logistical regression every point higher on the grit scale increases the odds of being a responder by 1.6 times. This set of finding helps us to understand what's going on. This means that our complete case respondents tended to be grittier. That's not terribly surprising. It does point out, though, that our complete case respondents aren't telling the story for the full study sample on this outcome. They are somehow different, perhaps a lower-risk group. So we're going to try to address this.

Step 3: Calculate nonresponse weights

High-level summary of Step 3

Step two really tells us a story about who the complete case sample is relative to the full sample? What things differentiate the complete case sample respondents from everyone else? Let's keep going. Knowing how our respondent's sample is different is kind of unsatisfying. Now we actually have information about why our sample is not representative. And the point now is there anything that we can do to help address this problem? And the answer is, yes. We can take what we learn from Step 2; we can create weights that we can use to improve our complete case analysis. And then we can make our complete case respondents better represent the full study sample by incorporating these weights. This is really going to address the key limitation that we've been worried about.

Nonresponse weights

Nonresponse weights are a way to adjust the data that we have in hand to compensate for loss of data. Essentially it gives some sample members more or less weight in computed statistics based on whether or not the observed individuals are good matches for the target sample. Where do we get the probability of being a respondent? Well, the good news is we can get that from that logit or probit model we did in Step 2. Remember, we try to predict whether a person was a complete case respondent, based on a bunch of demographic, and other key variables of interest. Notably, the baseline measure of the outcome. So Step 2 is really focused on the beta coefficients. What we're going to do here is ask our stats program to output the predicted probability for each individual in our data set. It's essentially the betas times each observation’s observed Xs. So every stats package is going to be able to do this. It's not a problem.

Illustrative SAS Code and output

[Displays SAS code and output; see following table.

```
Proc logistic data=my data;
Model STI_CC (event=’1’)= male Hispanic age
```

```
STI_knowledge_pre_GRIT/link=logit stb;
Output out=psdata
Predicted=P_ST_CC
Run;]
```

StudyID	STI_CC	P_STI_CC
101	1	.67
102	1	0.87
103	0	.42
104	1	.89
105	1	.91
106	0	.51
107	1	.49

Here's another code and output slide. It's actually the same code that I showed from Step 2, but with one small difference. The highlighted text is supposed to be the P STI CC stuff in the in the syntax on the left. On that is the predicted probability of being a complete case sample member. You can see the predicted probability is the new last column on the illustrative data set to the right. We have some folks who have high probabilities of being a complete case respondent. Those are, for example, observations 104 and 105 have pretty high probabilities. Some observations have pretty low probabilities; observation 103 and again your stats package, even if you're not using SAS can produce this.

Calculate nonresponse weights, and rescale

[Displays code and output table, see following table]

StudyID	STI_CC	P_STI_CC	STI_weight_raw	STI_weight_Rescaled
101	1	.67	1.49	1.08
102	1	0.87	1.15	.83
103	0	.42		
104	1	.89	1.12	.81
105	1	.91	1.10	.80
106	0	.51		
107	1	.49	2.04	1.48
Total	5		6.91	5

As shown here, the nonresponse weights are the inverse of the probability of being in the complete case sample. Someone who is likely to be a complete case sample member, for example, they had a high probability of being a respondent who tends to get lower weight than somebody who was a respondent, but was unlikely to do so. All we're going to need to do is take the inverse of the predicted probability. The weight is just one divided by the predicted probability. And you can see that as one of the columns in the slideshow. I'm going to ignore the weight for the observations that are not complete case sample members. They're not going to be included in the analysis. We're just going to ignore them for the weighting purposes. And we've got just empty cells for them.

One thing that I'd like for folks to look at is that we have a total of five complete case sample members. If you add up all of the ones in the STI underscore CC column. We add to a total of five but when I sum up the raw weights in this table, I have a bigger number; in bold, you can see the 6.91. That's at the bottom of the table. So this kind of helps me to understand that I can't use the raw weights in my analysis. It would make our sample seem bigger than it should be to our statistics package. We would get artificially low standard errors and artificially low *p*-values. All I'm going to do is rescale. In that final column, I'm going to multiply each weight by five, that's a true number of complete case sample members ,and divide by 6.91, that's the sum of the raw weights. And so that last row shows, actually that last column shows this. The sum of the new weights adds back up to five. This means that different folks are going to contribute more to the analysis than others. For example, observation 107 that individual is going to carry the greatest weight. Observation 105 with a weight of 0.80 is going to carry the least weight. You would end up doing the same type of analysis for your cases.

Incorporate the nonresponse weight in a revised version of the pre-post analysis

What would we do now that we have all of these weights? Well, we're going to produce exactly the same pre-post statistics that we previously reported. We will report baseline and follow-up means and standard deviations. We'll talk about a difference in means in terms of raw units, as well standardized units. We'll get a *p*-value. The big difference is that previously all of our observations had the same weight. They all counted the same. Now, we're going to weigh our observations by that nonresponse weight we just computed. So all of the computed stats that we're going to come up with are now going to take into account this nonresponse weight to make our reported statistics better representative of the target sample. This is really a marked improvement over what we did previously.

Illustrative SAS Code and output

[Displays two types of paired t-test codes in SAS:

```
Proc ttest data=psdata;
Where STI_CC = 1;
Paired STI_knowledge_pre *
Run;
```

N	Mean	Std Dev	Std Err	Min	Max
365	10.02	6.99	0.366	-9	30

```
proc ttest data=psdata;
Where STI_CC=1;
paired STI_knowledge pre *
weight STI_weight_rescaled;
    Run;
```

N	Mean	Std Dev	Std Err	Min	Max
365	9.95	7.34	0.366	-9	30

I'm going to show how easy it is to add weights to your analysis. I'm just going to compare two blocks of paired t-test code in SAS. The highlighted difference kind of indicates that we've added weights to our analysis. When you do this, you're going to get new statistics. In a paired t-test, the key thing is that we're looking at is the change from pre- to post-. You can see a different set of numbers in the two panels. For example, the mean on the left is 10.02; the mean on the right is 9.95. Other statistics are also going to be different. We're going to have different standard deviations, different standard errors; different *p*-values, but those aren't shown in this snip. But the key point here is it's very easy to add weights to your data and add weights to your analysis.

Step 4 (Bonus!): Show that your nonresponse weights improved representability

As a final bonus step, we can showcase how our weighted analysis checks boxes of credibility in terms of improving the representativeness of our results relative to a standard unweighted complete case pre-post analysis. What we're going to do in this final step is convince our audience that our nonresponse weights helped address the problem of our complete case sample not being fully representative of the full sample. We're going to show how those nonresponse weights helped us to recover numbers that were closer to the full population averages. It's really an exercise in doing three calculations for each variable of interest. We can report pre and post means for each variable of interest using three different approaches.

First, we'll estimate these pre and post means using all observed data. For example, the true means as best we know it. Second, we can estimate pre and post means for the complete case sample without weights. This is like the crude complete case analysis where all observations get equal weight. And finally, we can estimate pre and post means for the complete case sample after applying these nonresponse weights. This is hopefully going to get our analysis looking more like the true sample means from number one. Again, what we're hoping for here is that the results that we're going to observe from analysis number three, it's going to look more like number one, the true means than the analysis that we see in number two, the crude pre-post analysis that occurs without weighting. If we see improvement it helps us feel like our weights are doing a good job of improving the representativeness of our pre-post analysis.

Illustrative example: STI knowledge scores with and without weights

[Displays exam STI knowledge scores, see following table]

	Average at pre-test	Average at post-test
1.Based on all observed data	61.3	79.7
2.Complete case sample average, without weights	60	80
3.Complete case sample average, after weighting	60.5	79.9

Here's an illustrative example, again looking at STI Knowledge scores at pre and post. We've got three rows to our table. Row one shows the scores in pre and post using all observed data. We've got a 61.3 at pre and a 79.7 at post. These are actually the same numbers from the start of the presentation. That's 61.3, it includes two types of respondents. Those are folks with both pre-test data only, as well as our complete case sample members. This is a pretty rough guess for what the true population averages at pre because we're using all observed data. The same story for the 79.70 post. It's based on folks who are post-only respondents and our complete case sample members. It's as good guess at the true population averages we're going to get. The second row is our unweighted crude complete case analysis. We're purging from our estimates the post-only respondents and the post-only respondents who contributed to the average that was found in number one. We're focusing solely on the complete case sample and we're treating all observations with equal weight.

The third row is the complete case analysis after we've applied weights. What's nice here and why it’s shown in green is that the observed averages has been pulled closer to the population averages in row number one relative to the crude results that were shown in number two. This movement helps us feel like our complete case analysis is attempting to better represent the true population averages and thus that changes in scores are more appropriate to interpret for the full population, not just for the complete case sample. And again, we're doing the analysis that focuses on within-individual change and not pulling in composition changes. So it's the right thing to do. We sleep well knowing that we're talking about within-individual change rather than sample composition change. Doing this work helps close the loop on the nonresponse analysis. And the weighting work to help make the case more of that weighting—the complete case data makes the complete case sample more fully representative of the full target population.

This information in this third row, this is your story. This is the information that you should be using to report out as your benchmark outcome change for your pre-post analysis. This is the information that you should be using for talking about whether your logic model has been verified or validated. This is the information to use to as an upper-bound estimate for your power calculations. This is your story. You can always supplement this analysis with the unweighted results as a sensitivity result in case you don't fully trust your weights. But the key thing is that this is the message to report. You'd be able to say something like STI Knowledge improved from 60.5 at baseline to 79.9, an improvement of 19.4 points. That's the key message that you would want to report out to your readers.

Summary of key takeaways

Alright, let's quickly recap so there's time for questions. The first part of this webinar talked about the benefit of a single group pre-post outcome study. These types of studies quantify the extent to which participants' outcomes change, ideally improve over time. The big limitation is that they don't come with a counterfactual. What would have happened in the absence of the program? While we have a measure of how outcomes changed, we can't attribute that change in outcomes to the program we're implementing.

That being said, they helped establish some foundational evidence to set the stage for an impact or an effectiveness evaluation. First, you'll have some evidence that outcomes are trending in the right direction. That's important to show, if you want to do a rigorous impact evaluation. Hopefully, you'll be able to see that the outcomes in your logic model are improving, especially those outcomes that are the most proximal to the intervention improving the most. Again, this helps you to validate your logic model or your theory of change.

Second, these findings help give you a rough sense of the outcome change magnitude. I argued here that the pre-post change can give you an upper-bound estimate of the impact you might expect to see in an effectiveness evaluation. It's the upper bound because we expect that the control group will also experience some positive growth or maturation from relative to the business-as-usual program is being offered. So this information is helpful for assessing if your future evaluation is going to be sufficiently powered.

Best practices for pre-post analyses

And the latter half or two-thirds of the presentation focused on going beyond the basics with a pre-post analysis. I've argued to not be satisfied with the basic complete case analysis reporting on changes in outcome over time with descriptive and inferential results. Yes, it's important to report those statistics, but I'm encouraging you to go beyond those basics. Do report a response rate calculation to understand what item nonresponse and the profiles of your respondents look like. Ideally, we want to see that we have a high proportion of complete case respondents. Do a nonresponse analysis to understand the things that help differentiate your complete case respondents from everyone else. Don't be surprised when your complete case respondents tend to be a lower-risk population. And therefore, that you worry about generalizing findings from them. Do estimate weights as an easy add-on to the nonresponse analysis. And then plan on building those weights into your descriptive and inferential tests of how participants are changing over time. This is the easiest way to address the standard limitation that folks raise against a pre-post analysis, that your complete case isn't representative of the full study population.

And do that final analysis to show how the utilization of the nonresponse weights substantively improves the representability of the findings. It's a great way to close the loop and showcase the improved credibility of the findings. This is a lot of information, but hopefully this was helpful for everyone planning on conducting and reporting results from a post analysis as part of your final report and potentially your journal articles. As we talked about early on, the webinar recording will be made available in the future and also we're working on a written product to accompany this. Keep an eye out for that in the future. I think at this point, we're going to shift gears to Q&A, we will potentially see if there are any questions that were submitted during the presentation. And if not, please do submit some questions now. And I think that's it. Diana, do you want to take it from here?

Questions?

[Displays Russell Cole’s contact information: rcol@mathematica-mpr.com]

Diana McCallum: That was great. Thanks Russ. Sure. So as Russ mentioned, definitely go to the Q&A widget. That's going to the row below and clicking on Q&A. And you can submit your questions that way. We are certainly able to take questions now, but as Russ mentioned, we're going to record the webinar and the slides will be available later. And you can also just think through what you've heard from the webinar and see if you have any questions that you want to raise with the Evaluation TA liaisons. And we can discuss them with you on a future call, if that's helpful.

We have about nine minutes. If folks have questions or if you're thinking about things, we'll give you a few moments to put those together and submit them.

And I'll just say, we don't have any questions right now. But if folks think of questions, feel free to again submit them. We won't hang on too much longer. I'll wait a few more moments in case there are questions that pop up that you want an immediate answer to. But we think you know how to reach us and you can certainly also email Russ. The email that's up on the screen. We'll give it at least another minute. I know some of you may be typing in your questions or you may have something quick that you want to run by, that’s why we’re on the webinar, so we'll just hold if anything comes up.

And I'll just add, if you are trying to submit a question and maybe having some sort of technical difficulty you can raise a comment or chat to the tech widget. Please let us know.

Diana McCallum: Right. Well, I’m not going to keep folks too much longer on Friday. I know we've given you a fair amount to think about and just that you all will let us know if you have questions as you're working on analyses and work with you when we meet in a few weeks to discuss feedback on your abstracts. If that's what you'd like to do. Otherwise, please send an email to Russ or any of the members of the Evaluation TA team right. So with that we're going to go ahead and close out the webinar. And everybody, thanks for participating. We hope you enjoy the rest of your Friday. Thank you.

Russell Cole: Thanks all.